CUNY SPS DATA 621 - CTG5 - HW4

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 24th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET_FLAG	car crash = 1, no car crash = 0	response
TARGET_AMT	car crash cost = >0, no car crash = 0	response
AGE	driver's age - very young/old tend to be risky	numerical predictor
BLUEBOOK	$ value of vehicle	numerical predictor
CAR_AGE	age of vehicle	numerical predictor
CAR_TYPE	type of car (6types)	categorical predictor
CAR_USE	usage of car (commercial/private)	categorical predictor
CLM_FREQ	number of claims past 5 years	numerical predictor
EDUCATION	max education level (5types)	categorical predictor
HOMEKIDS	number of children at home	numerical predictor
HOME_VAL	$ value of home - home owners tend to drive more responsibly	numerical predictor
INCOME	$ income - rich people tend to get into fewer crashes	numerical predictor
JOB	job category (8types, 1missing)- white collar jobs tend to be safer	categorical predictor
KIDSDRIV	number of driving children - teenagers likely get into crashes	numerical predictor
MSTATUS	maritial status - married people drive more safely	catogerical predictor
MVR_PTS	number of traffic tickets	numerical predictor
OLDCLAIM	$ total claims in the past 5 years	numerical predictor
PARENT1	single parent	categorical predictor
RED_CAR	a red car	categorical predictor
REVOKED	license revoked (past 7 years) - more risky driver	categorical predictor
SEX	gender - woman may have less crashes than man	categorical predictor
TIF	time in force - number of years being customer	numerical predictor
TRAVTIME	distance to work	numerical predictor
URBANCITY	urban/rural	categorical predictor
YOJ	years on job - the longer they stay more safe	numerical predictor

1 DATA EXPLORATION

In this assignment, we explore, analyze and model a dataset containing 8,161 rows and 25 columns. Of all 25 features, 14 are discrete and 11 are continuous. There are total 970 missing values out of 204,025 observations.

We will build a logistic and multiple linear regression that will determine the followings: • Predict the probability that a person will crash their car • Amount of money it will cost if the person does crash their car

We will be able to develop insurance rates based on a number of predictors such as income, age, distance to work, and how long they have been customers, etc.

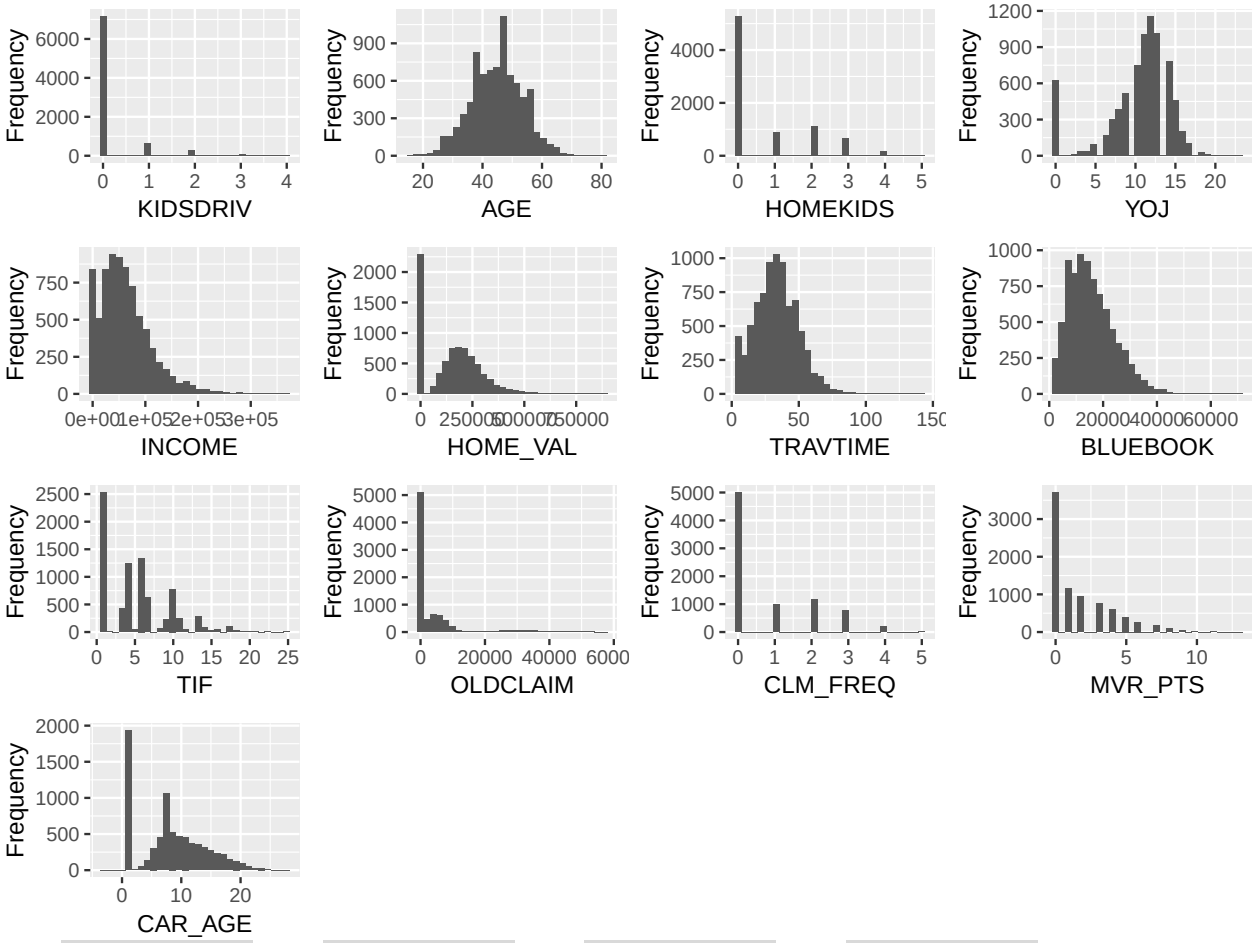
In the training dataset, there are 23 predictors and 2 response variables – one is binary value that indicates whether claim was made and the other is numerical value indicating the cost of claim.

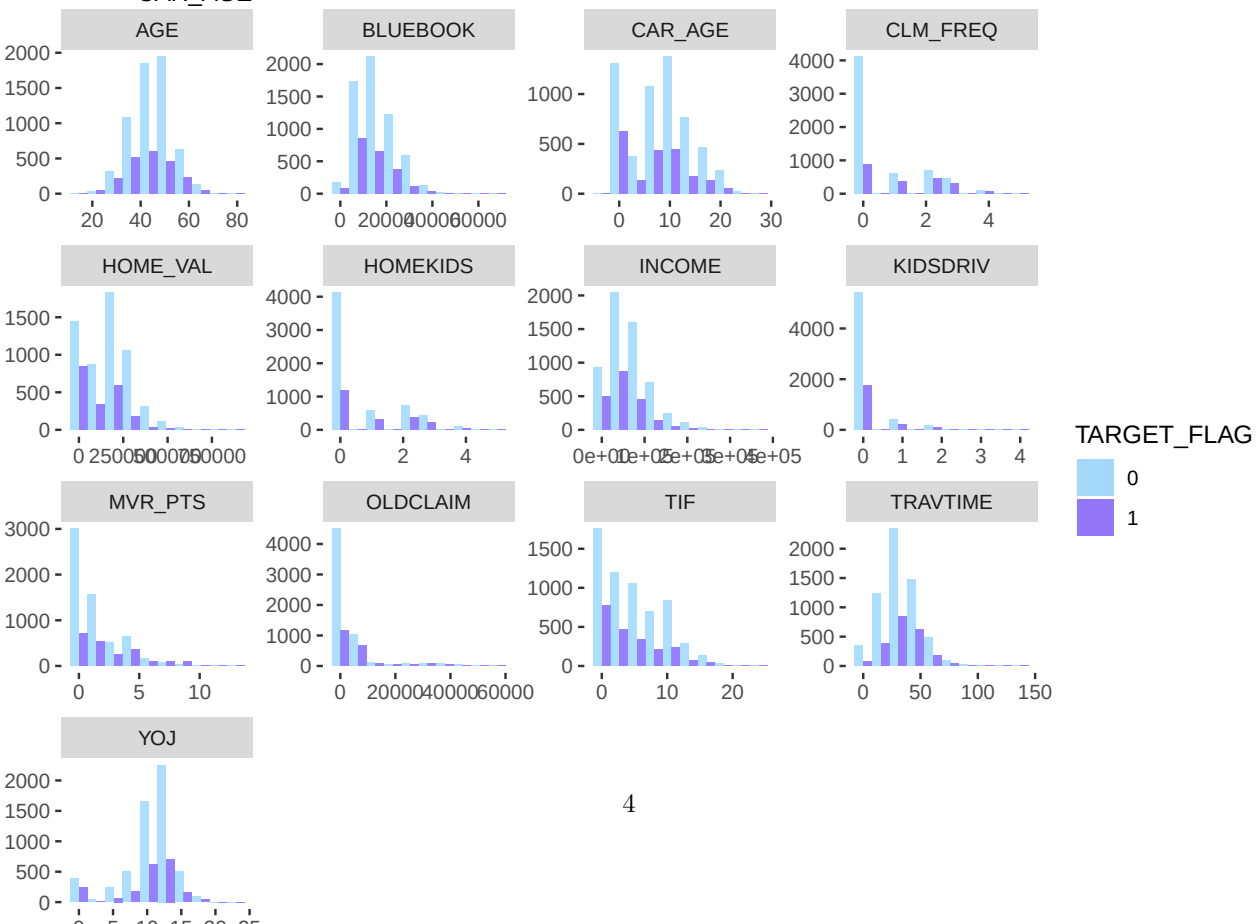
The response variable shows appropriate distribution in the training data. We confirm that for the number of target flags are 0 equals the target amount 0.

Table 2: Summary statistics

	n	mean	sd	median	min	max	skew	kurtosis
KIDSDRIV	8161	1.710575e-01	5.115341e-01	0	0	4	3.3518374	11.7801916
AGE	8155	4.479031e+01	8.627589e+00	45	16	81	-0.0289889	-0.0617020
HOMEKIDS	8161	7.212351e-01	1.116323e+00	0	0	5	1.3411271	0.6489915
YOJ	7707	1.049929e+01	4.092474e+00	11	0	23	-1.2029676	1.1773410
INCOME	7716	6.189809e+04	4.757268e+04	54028	0	367030	1.1863166	2.1290163
HOME_VAL	7697	1.548673e+05	1.291238e+05	161160	0	885282	0.4885950	-0.0160838
TRAVTIME	8161	3.348572e+01	1.590833e+01	33	5	142	0.4468174	0.6643331
BLUEBOOK	8161	1.570990e+04	8.419734e+03	14440	1500	69740	0.7942141	0.7913559
TIF	8161	5.351305e+00	4.146635e+00	4	1	25	0.8908120	0.4224940
OLDCLAIM	8161	4.037076e+03	8.777139e+03	0	0	57037	3.1190400	9.8606583
CLM_FREQ	8161	7.985541e-01	1.158453e+00	0	0	5	1.2087985	0.2842890
MVR_PTS	8161	1.695503e+00	2.147112e+00	1	0	13	1.3478403	1.3754900
CAR_AGE	7651	8.328323e+00	5.700742e+00	8	-3	28	0.2819531	-0.7489756

1.1 Summary Statistics





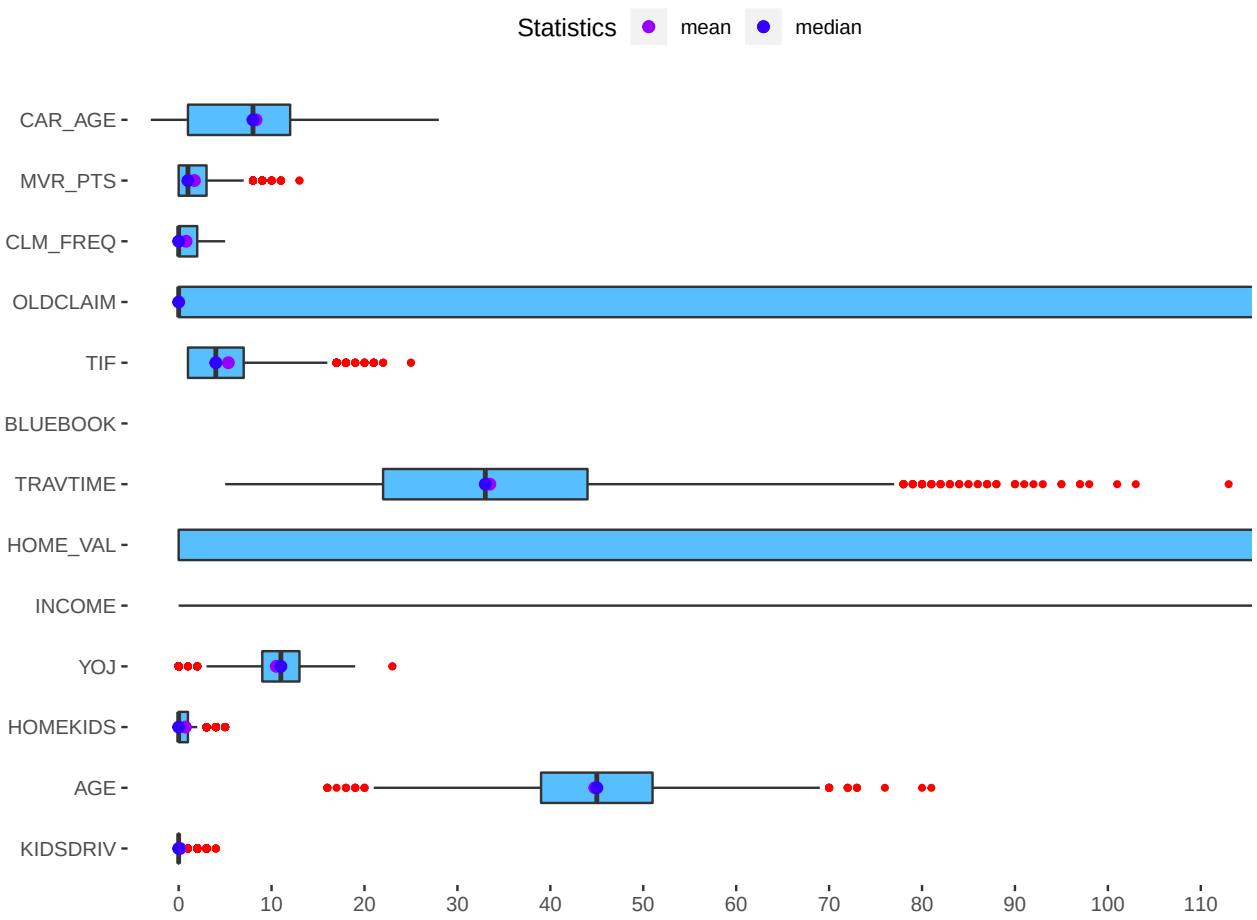


Figure 1: Outliers Boxplot

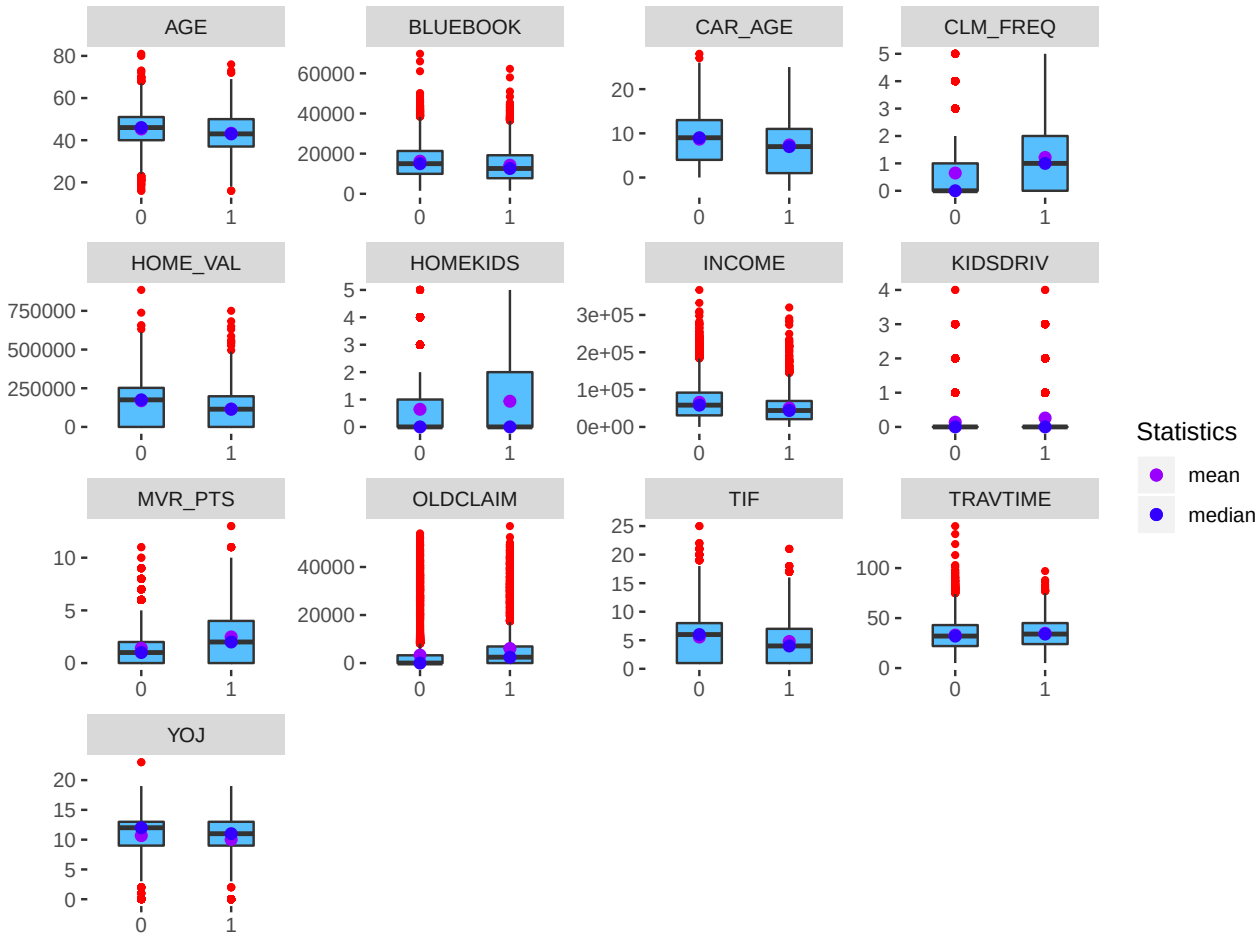
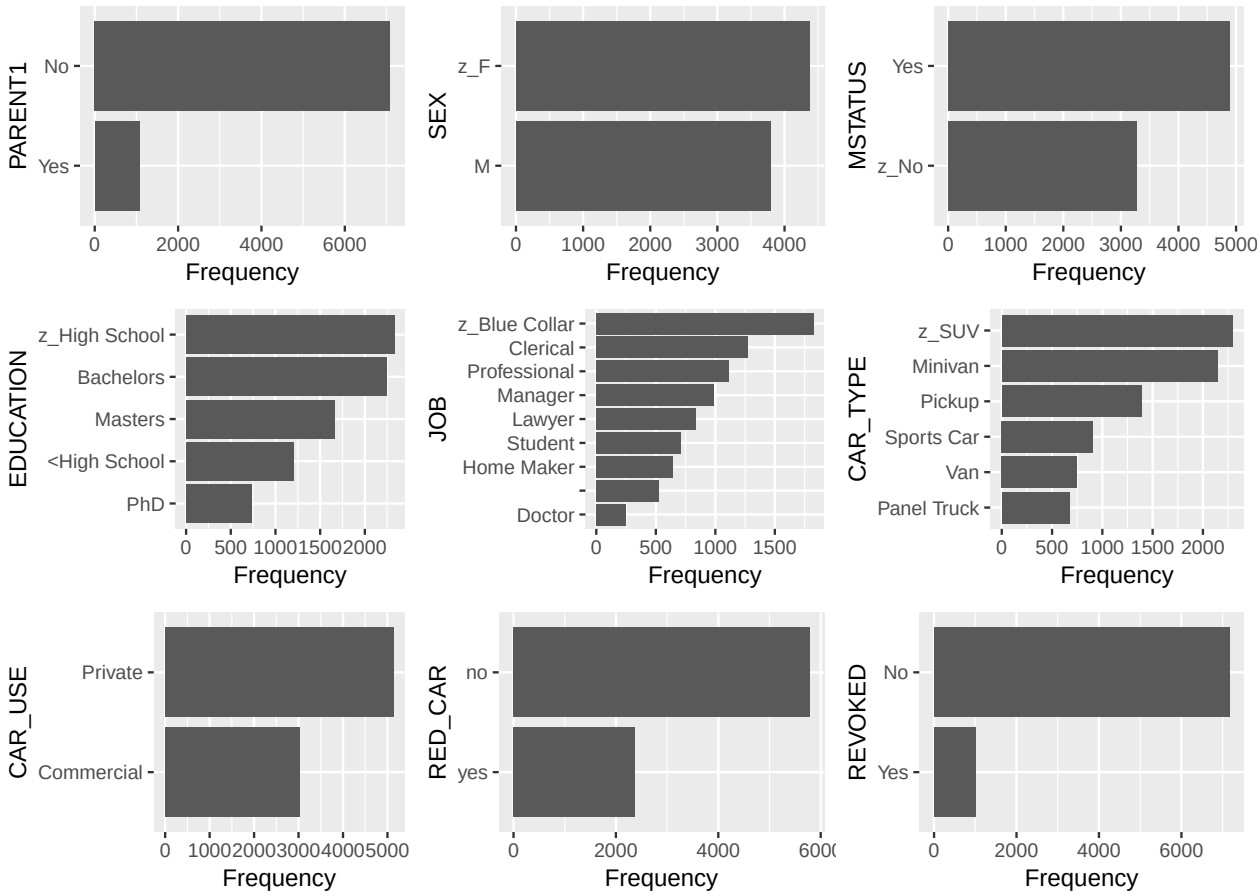
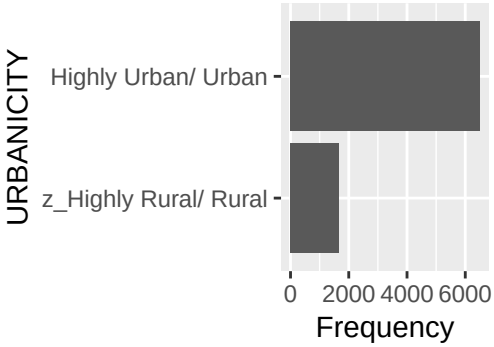


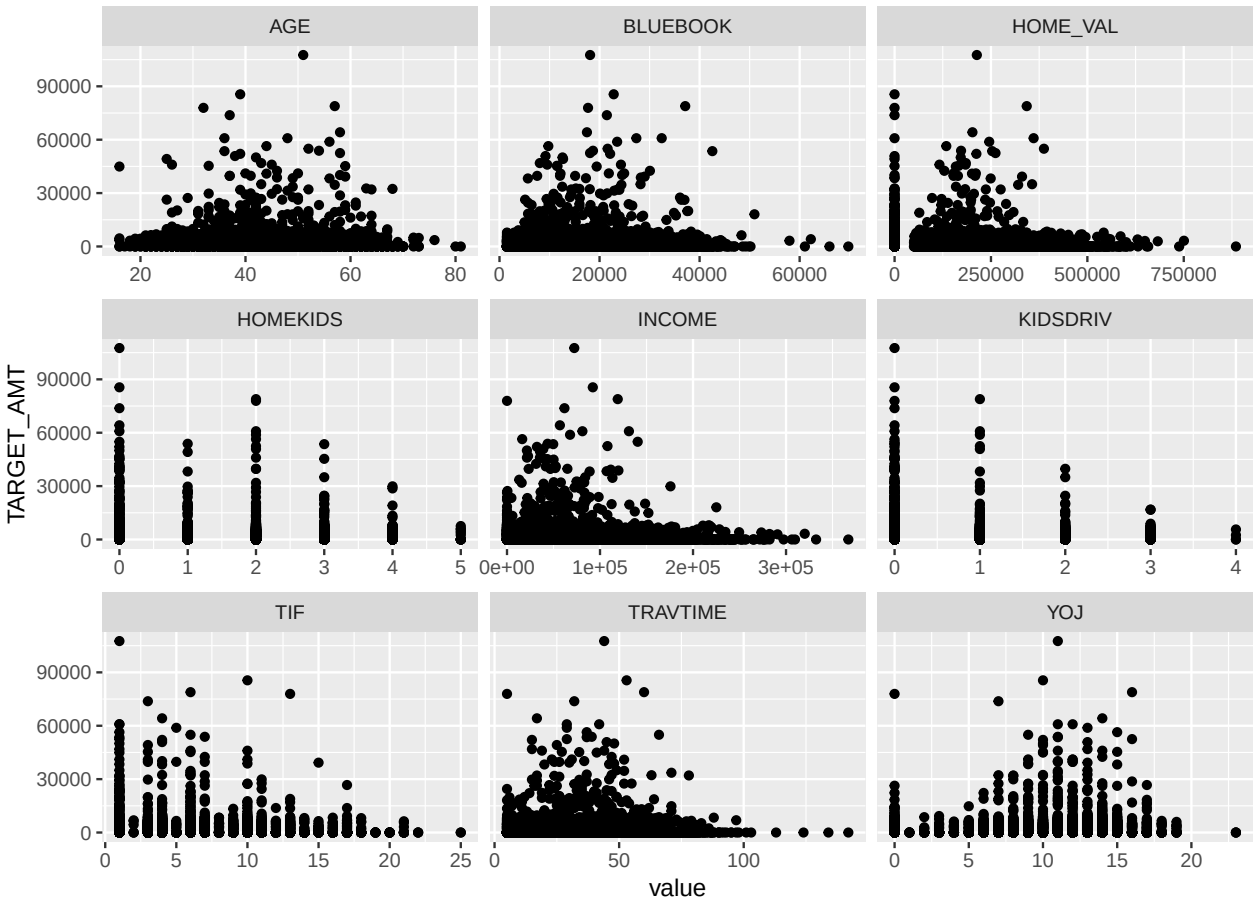
Figure 2: Linear relationship between each numeric predictors and the target

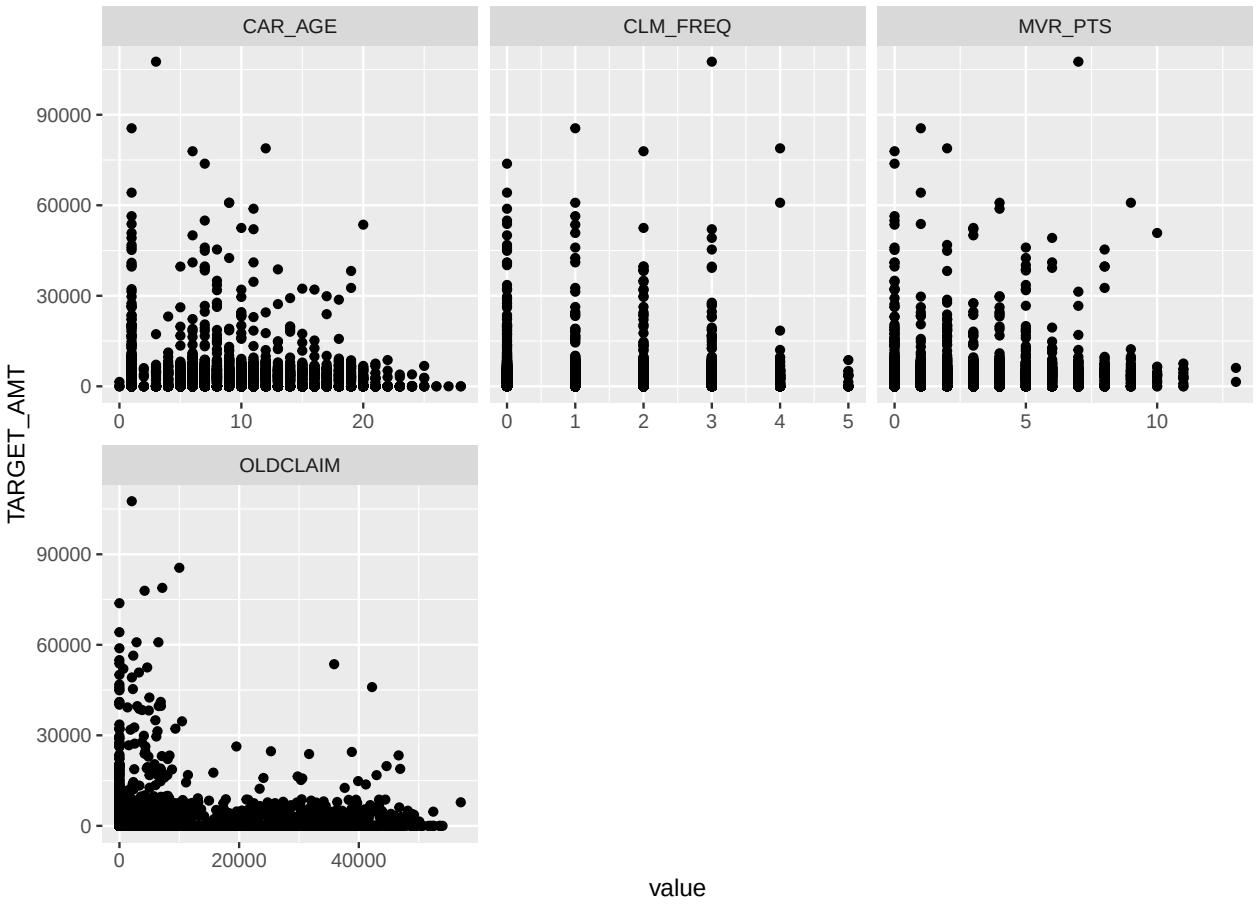

PARENT1
No
Yes
0 2000 4000 6000
Frequency
SEX
z_F
M
0 1000 2000 3000 4000
Frequency
MSTATUS
Yes
z_No
0 1000 2000 3000 4000 5000
Frequency
EDUCATION
z_High School
Bachelors
Masters
<High School
PhD
0 500 1000 1500 2000
Frequency
JOB
z_Blue Collar
Clerical
Professional
Manager
Lawyer
Student
Home Maker
-
Doctor
0 500 1000 1500
Frequency
CAR_TYPE
z_SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
0 500 1000 1500 2000
Frequency
CAR_USE
Private
Commercial
0 10002000300040005000
Frequency
RED_CAR
no
yes
0 2000 4000 6000
Frequency
REVOKED
No
Yes
0 2000 4000 6000
Frequency

Page 1


URBANICITY
Highly Urban/ Urban
z_Highly Rural/ Rural
0
2000
4000
6000
Frequency

Page 2


AGE
BLUEBOOK
HOME_VAL
HOMEKIDS
INCOME
KIDSDRIV
TIF
TRAVTIME
YOJ
TARGET_AMT
value
0
30000
60000
90000
20
40
60
80
0
20000
40000
60000
0
250000
500000
750000
0
1
2
3
4
5
0e+00
1e+05
2e+05
3e+05
0
5
10
15
20
25
0
50
100


CAR_AGE
CLM_FREQ
MVR_PTS
OLDCLAIM
TARGET_AMT
value
90000
60000
30000
0
0
10
20
0
1
2
3
4
5
0
5
10
20000
40000

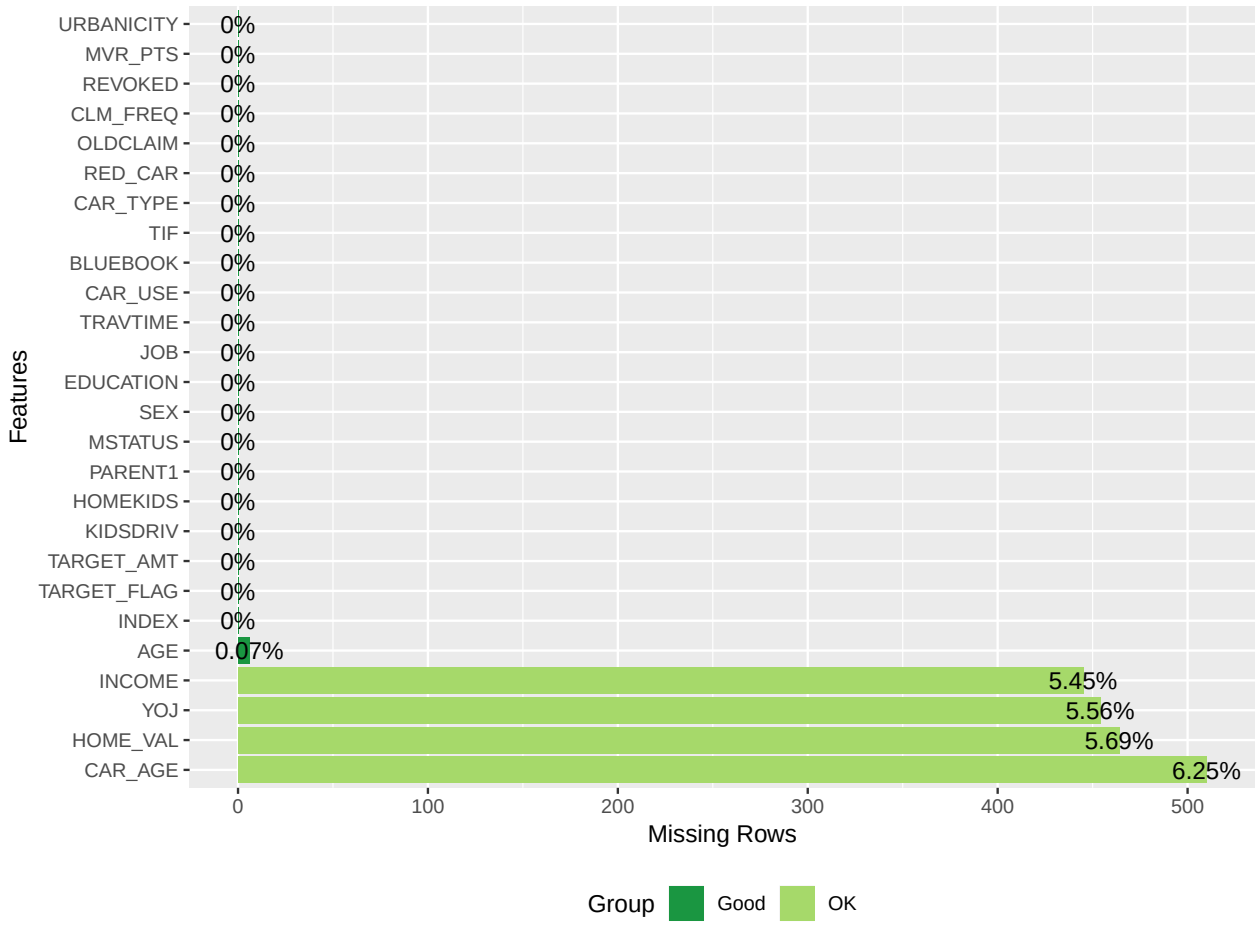


Figure 3: Missing data

There are a few missing data: `AGE`, `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`. Given the low proportion, it seems acceptable to impute the missing values.

2 DATA PREPARATION

2.1 Variable Desc

2.1.0.1 KIDSDRIV

`KIDSDRIV` is a discrete predictor with values ranging from 0 to 4. It shows heavy skewness with most cars having 0 kid drivers. Judging from the distribution, it appears that having kid driver results in higher probability of making a claim.

2.1.0.2 AGE

`AGE` presents driver's age and shows normal distribution, centered around 45. Looking at the boxplot of age, there is no difference between the claim made or not in distribution. Therefore, we can believe that `AGE` may not be helpful in determining the probability of making a claim.

2.1.0.3 HOMEKIDS

`HOMEKIDS` is a predictor describing number of children at home ranging from 0 to 5.

2.1.0.4 YOJ

`YOJ` is a predictor describing years on job. It is believed that people who stay at a job for a long time are usually more safe. YOJ shows normal distribution apart from those who are unemployed.

2.1.0.5 INCOME

`INCOME` is a heavily skewed predictor variable. The outliers should be treated.

2.1.0.6 HOME_VAL

`HOME_VAL` is a home value predictor variable. In theory, home owners tend to drive more responsibly. In the graph, we can see difference between the owners and renters.

2.1.0.7 TRAVTIME

`TRAVTIME` is a predictor variable describing the distance to work. Long drives to work usually suggest greater risk. However the graph shows fairly normal distribution and it may not be helpful determining the probability of making a claim.

2.1.0.8 BLUEBOOK

`BLUEBOOK` is a predictor variable describing the value of the car. The boxplot shows that the lower value of the car, the higher chances of making a claim. It is a possibility that the higher price cars are driven more carefully.

2.1.0.9 TIF

`TIF` describes how long the customer has been with the company, and the longer they have, the safer it may be. The plots show the safe drivers tend to stay safe.

2.1.0.10 OLDCLAIM

`OLDCLAIM` is a predictor describing the claims cost made in the past 5 years. We can see that it is very heavily skewed and that most people do not make claims.

2.1.0.11 CLM_FREQ

`CLM_FREQ` is a predictor that describes claim costs in the past 5 years. It seems that people who have made a claim in the past 5 years are highly likely to make another claim.

2.1.0.12 MVR_PTS

`MVR_PTS` is a predictor that describes motor vehicle record points. If you get lots of traffic tickets, you tend to get into more crash. It appears to be a highly significant variable as seen in boxplots.

2.1.0.13 CAR_AGE

`CAR_AGE` describes the vehicle age. There is one data point that shows the vehicle age is -3, this will be corrected to 0.

2.1.0.14 PARENT1

`PARENT1` describes single parent. This is factorized and renamed as `NumParents` to describe the number of parents.

2.1.0.15 SEX

`SEX` describes the gender of the driver. This is factorized and renamed as `MALE` to describe male as 1 and female as 0. It does not appear to be significant variable in the box plot.

2.1.0.16 MSTATUS

`MSTATUS` describes the martial status of the driver. It is believed that married people drive more safely. This variable has been factorized and renamed as `Single` to explain married as 0, not married as 1.

2.1.0.17 EDUCATION

`EDUCATION` describes the education level of the driver. It is factorized. It may be correlated with INCOME.

2.1.0.18 JOB

`JOB` describes the type of job the driver has. It is factorized. It may be correlated with INCOME. In theory white collar jobs tend to drive safer.

2.1.0.19 CAR_TYPE

`CAR_TYPE` describes type of car. It is factorized.

2.1.0.20 CAR_USE

`CAR_USE` describes how the car is used. Commercial vehicles are driven more and may increase probability of collision. It is factorized and renamed as `Commercial`. 0 means private.

2.1.0.21 RED_CAR

`RED_CAR` describes the color of the car is red. It is believed that red cars, especially sports cars are riskier. It is factorized.

2.1.0.22 REVOKED

`REVOKED` describes whether the license has revoked in the past 7 years. If it has revoked, it shows you are a risky driver. It is factorized. The boxplot shows the drivers who had lost their license are likely to be in accidents.

2.1.0.23 URBANICITY

`URBANICITY` describes whether driver lives in Urban area or Rural area. It is factorized and renamed as `URBAN`. 0 means rural.

2.2 Missing values

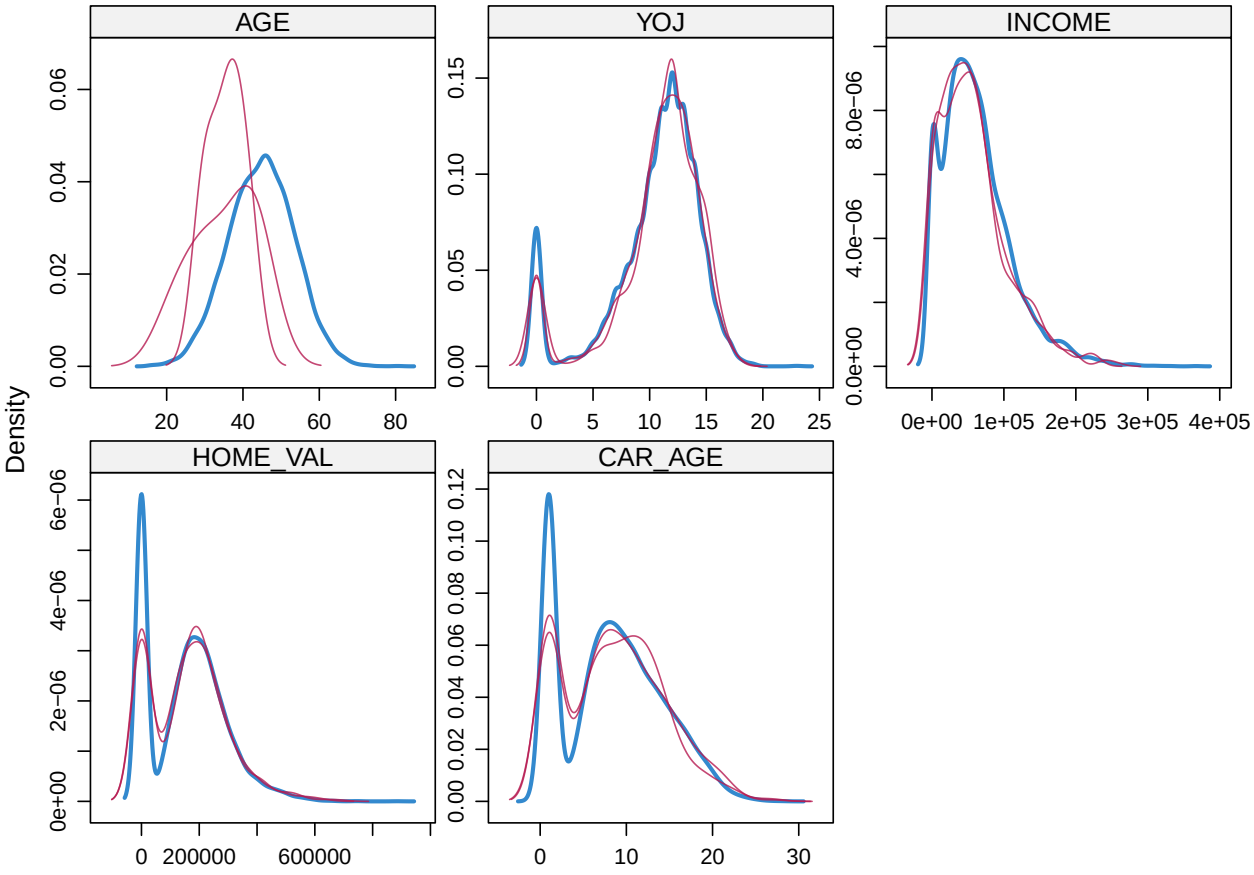
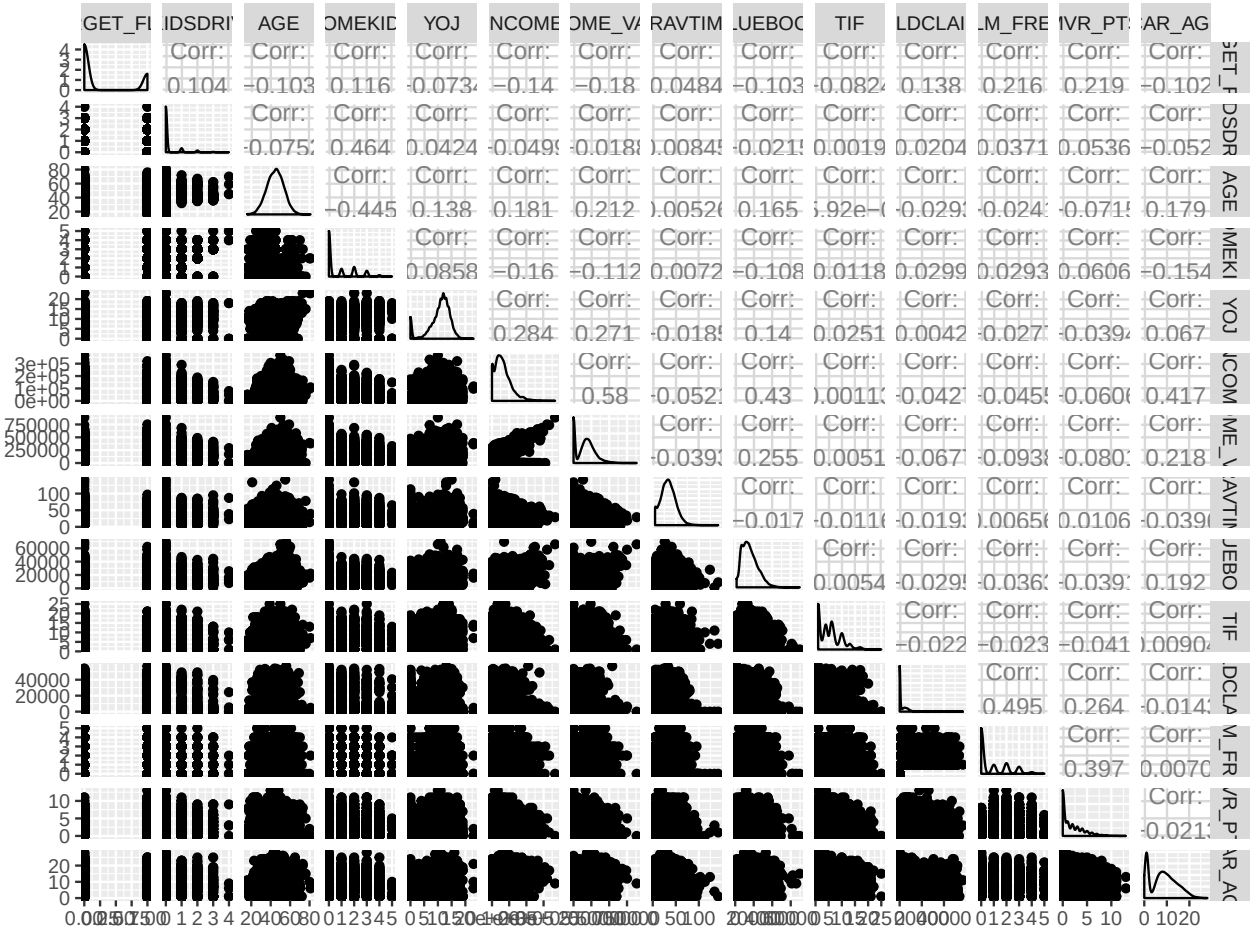


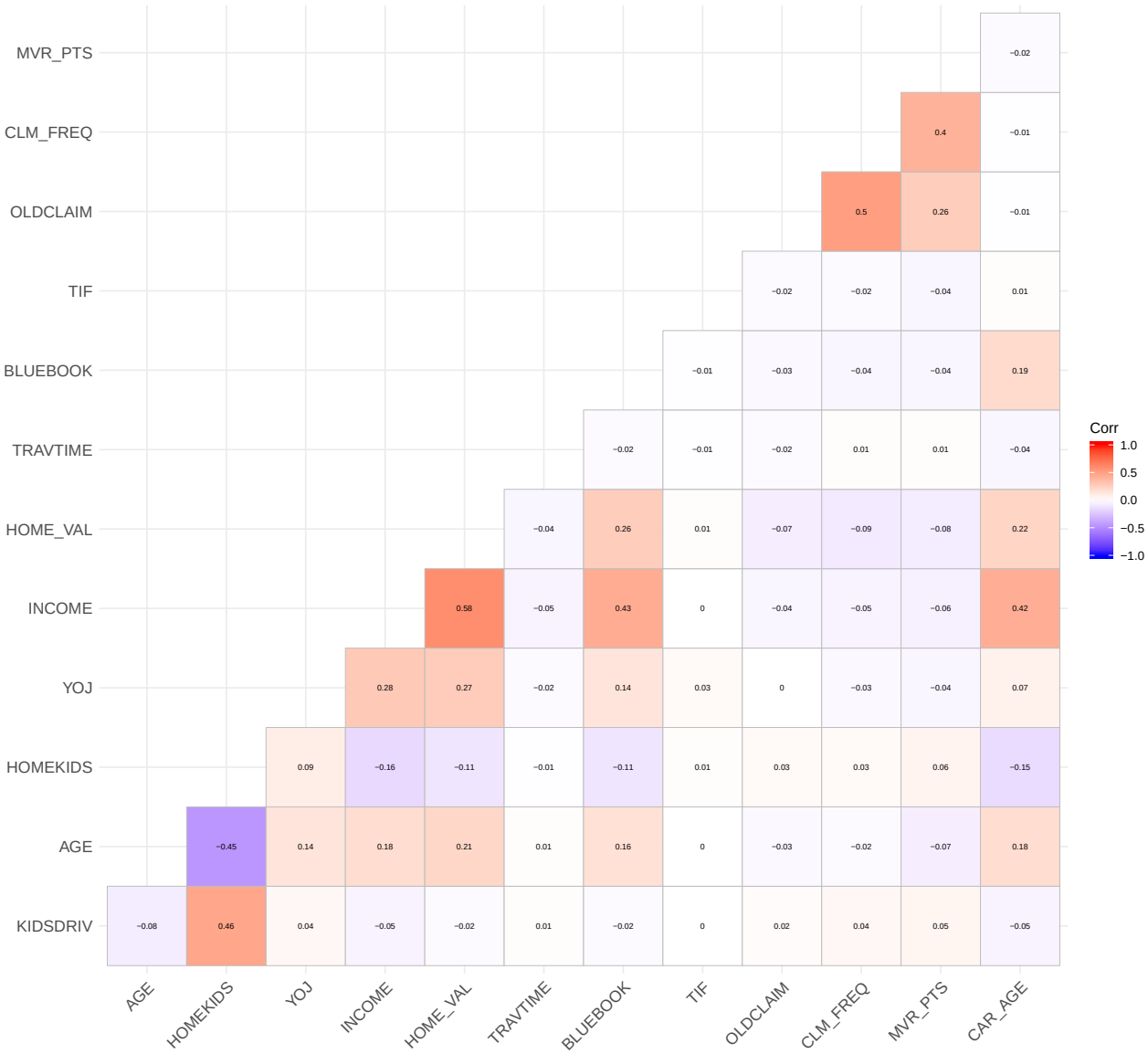
Figure 4: Difference between original and imputed data

We can see that excpet the `AGE`, the 4 variables roughly matches the existing distribution. We will use the 4 variables and impute `AGE` separately, using the median imputation.

Table 3: Correlation table

	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	
KIDSDRIV	1.00	-0.08	0.46	0.04	-0.05	-0.02	0.01	-0.02	
AGE	-0.08	1.00	-0.45	0.14	0.18	0.21	0.01	0.16	
HOMEKIDS	0.46	-0.45	1.00	0.09	-0.16	-0.11	-0.01	-0.11	
YOJ	0.04	0.14	0.09	1.00	0.28	0.27	-0.02	0.14	
INCOME	-0.05	0.18	-0.16	0.28	1.00	0.58	-0.05	0.43	
HOME_VAL	-0.02	0.21	-0.11	0.27	0.58	1.00	-0.04	0.26	
TRAVTIME	0.01	0.01	-0.01	-0.02	-0.05	-0.04	1.00	-0.02	-
BLUEBOOK	-0.02	0.16	-0.11	0.14	0.43	0.26	-0.02	1.00	-
TIF	0.00	0.00	0.01	0.03	0.00	0.01	-0.01	-0.01	
OLDCLAIM	0.02	-0.03	0.03	0.00	-0.04	-0.07	-0.02	-0.03	-
CLM_FREQ	0.04	-0.02	0.03	-0.03	-0.05	-0.09	0.01	-0.04	-
MVR_PTS	0.05	-0.07	0.06	-0.04	-0.06	-0.08	0.01	-0.04	-
CAR_AGE	-0.05	0.18	-0.15	0.07	0.42	0.22	-0.04	0.19	




MVR_PTS
CLM_FREQ
OLDCLAIM
TIF
BLUEBOOK
TRAVTIME
HOME_VAL
INCOME
YOJ
HOMEKIDS
AGE
KIDSDRIV
AGE
HOMEKIDS
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
CLM_FREQ
MVR_PTS
CAR_AGE
Corr
1.0
0.5
0.0
−0.5
−1.0

3 BUILD MODELS

3.1 Model 1

___ TARGET_FLAG ~ NumParents+ Male+ EDUCATION+ JOB+ CAR_TYPE+ RED_CAR+ REVOKED+ Urban+ Single+ Commercial ___

Model 1 only includes categorical variables as this will be easily interpretable and comprehedible when measuring the leading customers.

```
##
## Call:
## NULL
##
## Deviance Residuals:
##       Min          1Q      Median          3Q         Max
## -0.002447   0.000000   0.000000   0.000000   0.004127
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 426.1088  7635.6276   0.056    0.955
## TARGET_AMT                 1638.6589 14995.8480   0.109    0.913
## KIDSDRIV                      6.5241   344.8468   0.019    0.985
## AGE                          -3.0451   412.3513  -0.007    0.994
## HOMEKIDS                     -3.9888   521.9261  -0.008    0.994
## YOJ                          -6.4968   367.1233  -0.018    0.986
## INCOME                       -0.5194   473.2219  -0.001    0.999
## NumParents2                  -3.5776  1043.2917  -0.003    0.997
## HOME_VAL                      4.9674   724.7286   0.007    0.995
## MALEz_F                      -1.4700   861.1081  -0.002    0.999
## EDUCATIONBachelors           -7.0034  1089.8964  -0.006    0.995
## EDUCATIONMasters             -7.4923  1218.6958  -0.006    0.995
## EDUCATIONPhD                 -1.6586  1376.0731  -0.001    0.999
## `EDUCATIONz_High School`      0.5310   253.4587   0.002    0.998
## JOBClerical                  74.7169  6922.0482   0.011    0.991
## JOBDoctor                   -12.4408 27686.1495   0.000    1.000
## `JOBHome Maker`              49.5325  5286.6537   0.009    0.993
## JOBLawyer                    61.8129  6009.5946   0.010    0.992
## JOBManager                   66.1197  6414.6922   0.010    0.992
## JOBProfessional              65.2782  6900.2193   0.009    0.992
## JOBStudent                   55.5601  5401.5707   0.010    0.992
## `JOBz_Blue Collar`           87.7314  7943.1176   0.011    0.991
## TRAVTIME                     -1.4715   266.5086  -0.006    0.996
## BLUEBOOK                    -17.1393   829.2641  -0.021    0.984
## TIF                          -2.4442   275.0531  -0.009    0.993
## `CAR_TYPEPanel Truck`       -25.6470  5968.5770  -0.004    0.997
## CAR_TYPEPickup                2.9807   258.2479   0.012    0.991
## `CAR_TYPESports Car`          1.4375   574.4515   0.003    0.998
## CAR_TYPEVan                   3.5122  1018.2883   0.003    0.997
## CAR_TYPEz_SUV                -7.1926   901.2752  -0.008    0.994
## RED_CAR1                     -1.7358   253.5968  -0.007    0.995
## OLDCLAIM                     -0.6033   182.6953  -0.003    0.997
## CLM_FREQ                      5.1415   347.4286   0.015    0.988
## REVOKED1                      1.5849   270.9205   0.006    0.995
```

```
## MVR_PTS                     -0.5670   191.7749  -0.003    0.998
## CAR_AGE                      6.5615   256.4839   0.026    0.980
## Male1                            NA         NA      NA       NA
## Urban1                       4.6285   275.9668   0.017    0.987
## Single1                      1.9760   325.2707   0.006    0.995
## Commercial1                  1.8383   424.6487   0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.4180e+03  on 8160  degrees of freedom
## Residual deviance: 9.4642e-05  on 8122  degrees of freedom
## AIC: 78
##
## Number of Fisher Scoring iterations: 25
```

Df	Deviance	AIC
	7.75e+03	7.8e+03
1	7.8e+03	7.85e+03
1	7.76e+03	7.8e+03
4	7.79e+03	7.83e+03
8	7.88e+03	7.91e+03
5	7.9e+03	7.94e+03
1	7.75e+03	7.8e+03
1	7.85e+03	7.89e+03
1	8.53e+03	8.57e+03
1	7.82e+03	7.86e+03
1	7.83e+03	7.88e+03

```
##
## Call:  glm(formula = TARGET_FLAG ~ NumParents + Male + EDUCATION + JOB +
##     CAR_TYPE + REVOKED + Urban + Single + Commercial, family = "binomial",
##     data = train.disc.a)
##
## Coefficients:
##             (Intercept)             NumParents2                   Male1
##                -4.64408                 0.64262                 0.27129
##     EDUCATIONBachelors        EDUCATIONMasters             EDUCATIONPhD
##                -0.51381                -0.46903                 -0.55245
## EDUCATIONz_High School             JOBClerical               JOBDoctor
##                -0.04576                 0.66430                 -0.38335
##          JOBHome Maker               JOBLawyer              JOBManager
##                 0.74914                 0.14770                 -0.54325
##        JOBProfessional              JOBStudent        JOBz_Blue Collar
##                 0.25775                 0.81170                  0.45585
##    CAR_TYPEPanel Truck          CAR_TYPEPickup      CAR_TYPESports Car
##                 0.21238                 0.61370                  1.26512
##            CAR_TYPEVan           CAR_TYPEz_SUV                REVOKED1
```

```
##                0.45141                 0.98939                 0.76626
##                 Urban1                 Single1              Commercial1
##                2.40006                 0.53483                 0.77750
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8137 Residual
## Null Deviance:       9418
## Residual Deviance: 7751  AIC: 7799
```

`AIC` suggests that `RED_CAR` to be removed.

	x
TARGET_AMT	1.834984e+12
KIDSDRIV	9.703815e+08
AGE	1.387474e+09
HOMEKIDS	2.222840e+09
YOJ	1.099801e+09
INCOME	1.827342e+09
NumParents2	8.881814e+09
HOME_VAL	4.285889e+09
MALEz_F	6.050698e+09
EDUCATIONBachelors	9.693054e+09
EDUCATIONMasters	1.211939e+10
EDUCATIONPhD	1.545159e+10
‘EDUCATIONz_High School‘	5.242092e+08
JOBClerical	3.909844e+11
JOBDoctor	6.254827e+12
‘JOBHome Maker‘	2.280614e+11
JOBLawyer	2.947003e+11
JOBManager	3.357699e+11
JOBProfessional	3.885223e+11
JOBStudent	2.380840e+11
‘JOBz_Blue Collar‘	5.148398e+11
TRAVTIME	5.795791e+08
BLUEBOOK	5.611460e+09
TIF	6.173384e+08
‘CAR_TYPEPanel Truck‘	2.906911e+11
CAR_TYPEPickup	5.442065e+08
‘CAR_TYPESports Car‘	2.692755e+09
CAR_TYPEVan	8.461194e+09
CAR_TYPEz_SUV	6.628343e+09
RED_CAR1	5.247804e+08
OLDCLAIM	2.723609e+08
CLM_FREQ	9.849662e+08
REVOKED1	5.989268e+08
MVR_PTS	3.001054e+08
CAR_AGE	5.367975e+08
Male1	6.214466e+08
Urban1	8.633365e+08
Single1	1.471464e+09
Commercial1	1.834984e+12

3.2 Model 2

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5819  -0.7137  -0.3967   0.6205   3.1503
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.431367   0.034958 -40.945  < 2e-16 ***
## KIDSDRIV                  0.196827   0.031299   6.289 3.20e-10 ***
## AGE                      -0.003784   0.034716  -0.109 0.913202
## HOMEKIDS                  0.060313   0.041469   1.454 0.145834
## YOJ                      -0.064589   0.034506  -1.872 0.061234 .
## INCOME                   -0.163171   0.053494  -3.050 0.002287 **
## HOME_VAL                 -0.167865   0.044938  -3.735 0.000187 ***
## TRAVTIME                  0.231699   0.029967   7.732 1.06e-14 ***
## BLUEBOOK                 -0.173895   0.044370  -3.919 8.88e-05 ***
## TIF                      -0.229276   0.030464  -7.526 5.22e-14 ***
## OLDCLAIM                 -0.121408   0.034353  -3.534 0.000409 ***
## CLM_FREQ                  0.225534   0.033077   6.819 9.20e-12 ***
## MVR_PTS                   0.243287   0.029243   8.319  < 2e-16 ***
## CAR_AGE                  -0.017614   0.043194  -0.408 0.683438
## NumParents2               0.129724   0.037122   3.495 0.000475 ***
## Male1                     0.038547   0.049956   0.772 0.440347
## EDUCATIONBachelors       -0.163374   0.052009  -3.141 0.001682 **
## EDUCATIONMasters         -0.105068   0.072904  -1.441 0.149531
## EDUCATIONPhD             -0.034783   0.061908  -0.562 0.574215
## `EDUCATIONz_High School`  0.009900   0.042945   0.231 0.817685
## JOBClerical               0.143892   0.071593   2.010 0.044447 *
## JOBDoctor                -0.078940   0.045729  -1.726 0.084302 .
## `JOBHome Maker`           0.049171   0.056973   0.863 0.388112
## JOBLawyer                 0.029172   0.051451   0.567 0.570718
## JOBManager               -0.184713   0.056050  -3.296 0.000982 ***
## JOBProfessional           0.054808   0.061401   0.893 0.372058
## JOBStudent                0.045200   0.060983   0.741 0.458574
## `JOBz_Blue Collar`        0.125611   0.077451   1.622 0.104840
## `CAR_TYPEPanel Truck`     0.153626   0.044591   3.445 0.000571 ***
## CAR_TYPEPickup            0.209100   0.037864   5.522 3.35e-08 ***
## `CAR_TYPESports Car`      0.321552   0.040841   7.873 3.45e-15 ***
## CAR_TYPEVan               0.178608   0.036532   4.889 1.01e-06 ***
## CAR_TYPEz_SUV             0.345695   0.049997   6.914 4.70e-12 ***
## REVOKED1                  0.291478   0.029969   9.726  < 2e-16 ***
## Urban1                    0.966523   0.045523  21.231  < 2e-16 ***
## Single1                   0.236552   0.041889   5.647 1.63e-08 ***
## Commercial1               0.366000   0.044316   8.259  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7293.1  on 8124  degrees of freedom
## AIC: 7367.1
##
## Number of Fisher Scoring iterations: 5
```

Df	Deviance	AIC
	7.29e+03	7.37e+03
1	7.33e+03	7.4e+03
1	7.29e+03	7.37e+03
1	7.3e+03	7.37e+03
1	7.3e+03	7.37e+03
1	7.3e+03	7.37e+03
1	7.31e+03	7.38e+03
1	7.35e+03	7.43e+03
1	7.31e+03	7.38e+03
1	7.35e+03	7.42e+03
1	7.31e+03	7.38e+03
1	7.34e+03	7.41e+03
1	7.36e+03	7.43e+03
1	7.29e+03	7.37e+03
1	7.31e+03	7.38e+03
1	7.29e+03	7.37e+03
4	7.31e+03	7.38e+03
8	7.36e+03	7.41e+03
5	7.38e+03	7.45e+03
1	7.39e+03	7.46e+03
1	7.95e+03	8.02e+03
1	7.32e+03	7.4e+03
1	7.36e+03	7.43e+03

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME +
##     HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
##     MVR_PTS + NumParents + EDUCATION + JOB + CAR_TYPE + REVOKED +
##     Urban + Single + Commercial, family = "binomial", data = train)
##
## Coefficients:
##             (Intercept)                KIDSDRIV                HOMEKIDS
##              -4.071e+00               3.834e-01               5.413e-02
```

```
##                     YOJ                  INCOME               HOME_VAL
##             -1.580e-02              -3.464e-06             -1.290e-06
##               TRAVTIME                BLUEBOOK                    TIF
##              1.457e-02              -2.240e-05             -5.531e-02
##               OLDCLAIM                CLM_FREQ                 MVR_PTS
##             -1.383e-05              1.947e-01              1.133e-01
##            NumParents2      EDUCATIONBachelors        EDUCATIONMasters
##              3.827e-01              -3.824e-01             -2.947e-01
##           EDUCATIONPhD  EDUCATIONz_High School             JOBClerical
##             -1.576e-01              1.853e-02              3.941e-01
##              JOBDoctor           JOBHome Maker              JOBLawyer
##             -4.565e-01              1.667e-01              9.595e-02
##             JOBManager         JOBProfessional             JOBStudent
##             -5.665e-01              1.579e-01              1.561e-01
##       JOBz_Blue Collar     CAR_TYPEPanel Truck          CAR_TYPEPickup
##              3.006e-01              6.022e-01              5.556e-01
##     CAR_TYPESports Car             CAR_TYPEVan           CAR_TYPEz_SUV
##              9.679e-01              6.444e-01              7.147e-01
##               REVOKED1                 Urban1                Single1
##              8.900e-01              2.397e+00              4.831e-01
##            Commercial1
##              7.565e-01
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8127 Residual
## Null Deviance:       9418
## Residual Deviance: 7294  AIC: 7362
```

AIC suggest to remove AGE, CAR_AGE and Male

	x
KIDSDRIV	7.993589
AGE	9.834492
HOMEKIDS	14.032741
YOJ	9.716027
INCOME	23.351143
HOME_VAL	16.478599
TRAVTIME	7.327727
BLUEBOOK	16.064453
TIF	7.572833
OLDCLAIM	9.630043
CLM_FREQ	8.927504
MVR_PTS	6.978087
CAR_AGE	15.224372
NumParents2	11.244595
Male1	20.364411
EDUCATIONBachelors	22.072690
EDUCATIONMasters	43.370033
EDUCATIONPhD	31.274322
‘EDUCATIONz_High School‘	15.049065
JOBClerical	41.824816
JOBDoctor	17.063985
‘JOBHome Maker‘	26.487149
JOBLawyer	21.601284
JOBManager	25.635279
JOBProfessional	30.763987
JOBStudent	30.346433
‘JOBz_Blue Collar‘	48.948485
‘CAR_TYPEPanel Truck‘	16.224725
CAR_TYPEPickup	11.699148
‘CAR_TYPESports Car‘	13.610561
CAR_TYPEVan	10.889971
CAR_TYPEz_SUV	20.397524
REVOKED1	7.328760
Urban1	16.910525
Single1	14.318075
Commercial1	16.025632

3.3 Model 3

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5784  -0.7157  -0.3968   0.6254   3.1564
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.43133    0.03496 -40.943  < 2e-16 ***
## KIDSDRIV                  0.19611    0.03079   6.369 1.90e-10 ***
```

```
## HOMEKIDS                   0.06043    0.03846   1.571 0.116119
## YOJ                       -0.06468    0.03398  -1.903 0.056981 .
## INCOME                    -0.16469    0.05339  -3.085 0.002039 **
## HOME_VAL                  -0.16696    0.04480  -3.727 0.000194 ***
## TRAVTIME                   0.23177    0.02995   7.738 1.01e-14 ***
## BLUEBOOK                  -0.18861    0.03986  -4.732 2.22e-06 ***
## TIF                       -0.22934    0.03046  -7.530 5.09e-14 ***
## OLDCLAIM                  -0.12141    0.03435  -3.534 0.000409 ***
## CLM_FREQ                   0.22559    0.03306   6.822 8.95e-12 ***
## MVR_PTS                    0.24319    0.02923   8.320  < 2e-16 ***
## NumParents2                0.12954    0.03694   3.506 0.000454 ***
## EDUCATIONBachelors        -0.17070    0.04874  -3.503 0.000461 ***
## EDUCATIONMasters          -0.11859    0.06514  -1.821 0.068672 .
## EDUCATIONPhD              -0.04492    0.05759  -0.780 0.435313
## `EDUCATIONz_High School`   0.00837    0.04281   0.196 0.844977
## JOBClerical                0.14293    0.07156   1.997 0.045789 *
## JOBDoctor                 -0.07805    0.04568  -1.709 0.087516 .
## `JOBHome Maker`            0.04485    0.05666   0.792 0.428591
## JOBLawyer                  0.02908    0.05137   0.566 0.571306
## JOBManager                -0.18480    0.05600  -3.300 0.000968 ***
## JOBProfessional            0.05428    0.06139   0.884 0.376577
## JOBStudent                 0.04406    0.06095   0.723 0.469739
## `JOBz_Blue Collar`         0.12525    0.07743   1.618 0.105768
## `CAR_TYPEPanel Truck`      0.16599    0.04162   3.988 6.66e-05 ***
## CAR_TYPEPickup             0.20881    0.03784   5.519 3.41e-08 ***
## `CAR_TYPESports Car`       0.30424    0.03380   9.002  < 2e-16 ***
## CAR_TYPEVan                0.18616    0.03529   5.275 1.33e-07 ***
## CAR_TYPEz_SUV              0.32129    0.03867   8.308  < 2e-16 ***
## REVOKED1                   0.29185    0.02996   9.740  < 2e-16 ***
## Urban1                     0.96700    0.04552  21.244  < 2e-16 ***
## Single1                    0.23672    0.04188   5.652 1.59e-08 ***
## Commercial1                0.36550    0.04429   8.252  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7293.9  on 8127  degrees of freedom
## AIC: 7361.9
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME +
##     HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
##     MVR_PTS + NumParents + EDUCATION + JOB + CAR_TYPE + REVOKED +
##     Urban + Single + Commercial, family = "binomial", data = train)
##
## Coefficients:
##            (Intercept)               KIDSDRIV               HOMEKIDS
##             -4.071e+00              3.834e-01              5.413e-02
##                    YOJ                 INCOME               HOME_VAL
##             -1.580e-02             -3.464e-06             -1.290e-06
```

Df	Deviance	AIC
	7.29e+03	7.36e+03
1	7.33e+03	7.4e+03
1	7.3e+03	7.36e+03
1	7.3e+03	7.36e+03
1	7.3e+03	7.37e+03
1	7.31e+03	7.37e+03
1	7.35e+03	7.42e+03
1	7.32e+03	7.38e+03
1	7.35e+03	7.42e+03
1	7.31e+03	7.37e+03
1	7.34e+03	7.41e+03
1	7.36e+03	7.43e+03
1	7.31e+03	7.37e+03
4	7.32e+03	7.38e+03
8	7.36e+03	7.41e+03
5	7.4e+03	7.46e+03
1	7.39e+03	7.45e+03
1	7.95e+03	8.01e+03
1	7.33e+03	7.39e+03
1	7.36e+03	7.43e+03

```
##              TRAVTIME               BLUEBOOK                    TIF
##             1.457e-02             -2.240e-05             -5.531e-02
##              OLDCLAIM               CLM_FREQ                MVR_PTS
##             -1.383e-05              1.947e-01              1.133e-01
##           NumParents2     EDUCATIONBachelors       EDUCATIONMasters
##             3.827e-01             -3.824e-01             -2.947e-01
##          EDUCATIONPhD EDUCATIONz_High School           JOBClerical
##            -1.576e-01              1.853e-02              3.941e-01
##             JOBDoctor          JOBHome Maker             JOBLawyer
##            -4.565e-01              1.667e-01              9.595e-02
##            JOBManager        JOBProfessional             JOBStudent
##            -5.665e-01              1.579e-01              1.561e-01
##      JOBz_Blue Collar    CAR_TYPEPanel Truck         CAR_TYPEPickup
##             3.006e-01              6.022e-01              5.556e-01
##    CAR_TYPESports Car           CAR_TYPEVan         CAR_TYPEz_SUV
##             9.679e-01              6.444e-01              7.147e-01
##              REVOKED1                 Urban1                Single1
##             8.900e-01              2.397e+00              4.831e-01
##          Commercial1
```

```
##              7.565e-01
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8127 Residual
## Null Deviance:       9418
## Residual Deviance: 7294  AIC: 7362
```

	x
KIDSDRIV	7.736669
HOMEKIDS	12.069257
YOJ	9.421389
INCOME	23.260669
HOME_VAL	16.376158
TRAVTIME	7.320702
BLUEBOOK	12.964139
TIF	7.570256
OLDCLAIM	9.628115
CLM_FREQ	8.921550
MVR_PTS	6.971491
NumParents2	11.136841
EDUCATIONBachelors	19.380696
EDUCATIONMasters	34.624382
EDUCATIONPhD	27.060116
‘EDUCATIONz_High School‘	14.951849
JOBClerical	41.784863
JOBDoctor	17.027884
‘JOBHome Maker‘	26.193644
JOBLawyer	21.531651
JOBManager	25.594394
JOBProfessional	30.750101
JOBStudent	30.309948
‘JOBz_Blue Collar‘	48.927232
‘CAR_TYPEPanel Truck‘	14.136144
CAR_TYPEPickup	11.681883
‘CAR_TYPESports Car‘	9.321133
CAR_TYPEVan	10.163901
CAR_TYPEz_SUV	12.202862
REVOKED1	7.326779
Urban1	16.907074
Single1	14.313797
Commercial1	16.006842

3.4 Model 4

Observations	8161
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(32)$	2138.80
Pseudo-R^2 (Cragg-Uhler)	0.34
Pseudo-R^2 (McFadden)	0.23
AIC	7345.16
BIC	7576.39

	Est.	S.E.	z val.	p	VIF
(Intercept)	-1.12	0.79	-1.41	0.16	NA
KIDSDRIV	0.43	0.06	7.72	0.00	1.10
log(AGE)	-0.30	0.16	-1.95	0.05	1.26
YOJ	0.01	0.01	1.37	0.17	2.34
log(INCOME + 1e-14)	-0.02	0.00	-4.47	0.00	3.21
HOME_VAL	-0.00	0.00	-5.05	0.00	1.75
log(TRAVTIME)	0.41	0.05	7.93	0.00	1.03
log(BLUEBOOK)	-0.32	0.06	-5.82	0.00	1.48
TIF	-0.05	0.01	-7.31	0.00	1.01
log(OLDCLAIM + 1e-14)	0.01	0.00	6.29	0.00	1.26
MVR_PTS	0.10	0.01	7.02	0.00	1.24
NumParents2	0.40	0.10	3.97	0.00	1.60
EDUCATIONBachelors	-0.41	0.11	-3.81	0.00	7.46
EDUCATIONMasters	-0.34	0.16	-2.13	0.03	7.46
EDUCATIONPhD	-0.30	0.19	-1.57	0.12	7.46
EDUCATIONz_High School	0.03	0.09	0.27	0.79	7.46
JOBClerical	0.50	0.19	2.55	0.01	26.39
JOBDoctor	-0.39	0.27	-1.48	0.14	26.39
JOBHome Maker	0.12	0.21	0.57	0.57	26.39
JOBLawyer	0.17	0.17	1.02	0.31	26.39
JOBManager	-0.51	0.17	-3.01	0.00	26.39
JOBProfessional	0.22	0.18	1.26	0.21	26.39
JOBStudent	0.09	0.22	0.41	0.68	26.39
JOBz_Blue Collar	0.39	0.19	2.09	0.04	26.39
CAR_TYPEPanel Truck	0.53	0.14	3.69	0.00	2.32
CAR_TYPEPickup	0.58	0.10	5.82	0.00	2.32
CAR_TYPESports Car	0.96	0.11	8.91	0.00	2.32
CAR_TYPEVan	0.65	0.12	5.34	0.00	2.32
CAR_TYPEz_SUV	0.74	0.09	8.61	0.00	2.32
REVOKED1	0.71	0.08	8.87	0.00	1.01
Urban1	2.36	0.11	20.89	0.00	1.14
Single1	0.46	0.08	5.62	0.00	1.94
Commercial1	0.75	0.09	8.18	0.00	2.46

Standard errors: MLE

4 SELECT MODELS

5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining